DOCUMENT RESUME

ED 456 151 TM 033 214

AUTHOR	Williams, Kelly; Wise, Steven L.; West, Richard F.
TITLE	Multifaceted Measurement of Critical Thinking Skills in College Students.
PUB DATE	2001-04-00
NOTE	29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*College Students; *Critical Thinking; Higher Education; Measures (Individuals); Metacognition; *Skill Development; Student Attitudes; Test Construction; *Thinking Skills

ABSTRACT

This study was conducted to measure the development of critical thinking in college students through the creation of instruments to measure critical thinking skills. A Critical Thinking Assessment (CTA) was developed based on other such instruments and the literature. This instrument was tested in a pilot study involving 750 incoming university freshmen and a subsequent study of the revised instrument involving 730 sophomores. The CTA was designed to measure four skill areas: analysis, evaluation, inference, and interpretation. An instrument to measure student dispositions, a shortened version of the measure of actively open-minded thinking (AOT) of K. Stanovich and R. West (1997), was administered to the same students. The Ways We Think (WWT) metacognition evaluation was administered to 196 sophomores in February 2001. A Writing Rubric (CTWRITE) is also under development. In spite of the limitations of sample size and the possibility that the sample may not generalize beyond James Madison University, the data indicate that the CTA, the AOT, and the WWT measures demonstrate adequate reliability for making group-level inferences. The psychometric properties of these instruments are equal to, if not better, than other instruments measuring similar domains. (Contains 31 references, 4 figures, and 6 tables.) (SLD)

Multifaceted Measurement of Critical Thinking Skills In College Students

Kelly Williams, M.S.
Steven L. Wise, Ph.D.
Richard F. West, Ph.D.

James Madison University

Presented at the Annual Meeting of the National Council on Measurement in Education
Seattle, Washington, April 10-14

Introduction

The primary objective of this research project is to measure the development of critical thinking (CT) skills in college students. This poster presents the first step in the overall project; that is, identifying and/or creating reliable and valid instruments. The design entails the development of multiple assessments: a critical thinking skills test, a critical thinking dispositions inventory, a measure of metacognition, and rating of student argument presentation, both oral and written. These measures taken together address seven different facets within the critical thinking domain that we consider measurable. This is in accord with current commonly held conceptions of critical thinking both within academia and the business world (Facione, 1990; Delphi Report, 1990; Erwin, 2001). An exploration of the relationships among and between the different manifestations of skill in critical thinking is undertaken.

Literature Review

There is not one particular theorist who captures the eclectic framework within which this research is grounded. Rather this project progresses from a definition gleaned from studying a number of different critical thinking experts (notably Baron, 1998; Ennis, Millman, & Tomko, 1985; Erwin, 2001; Halpern, 1997; Kelley, 1988; Manktelow, 1999; Stanovich & West, 1999a; Stanovich & West, 2000). A central proposition of this project's working definition of critical thinking is that it results in judgments and decisions that optimize goal satisfaction and the accuracy of our representations of the objective world. In application, critical thinking entails taking in information and apprehending its relevance. These are critical components of rationality as studied by psychologists and philosophers.

While the focus here is on the initial measurement of critical thinking and the related tools of measurement, it is only one phase of a larger project. We suspect that course exposures in general education lay a foundation for the development of critical thinking. The teaching and acquisition of critical thinking skills may then continue throughout the college career in the different majors. To test the veracity of this hypothesis we must first establish valid and reliable means of measuring critical thinking.

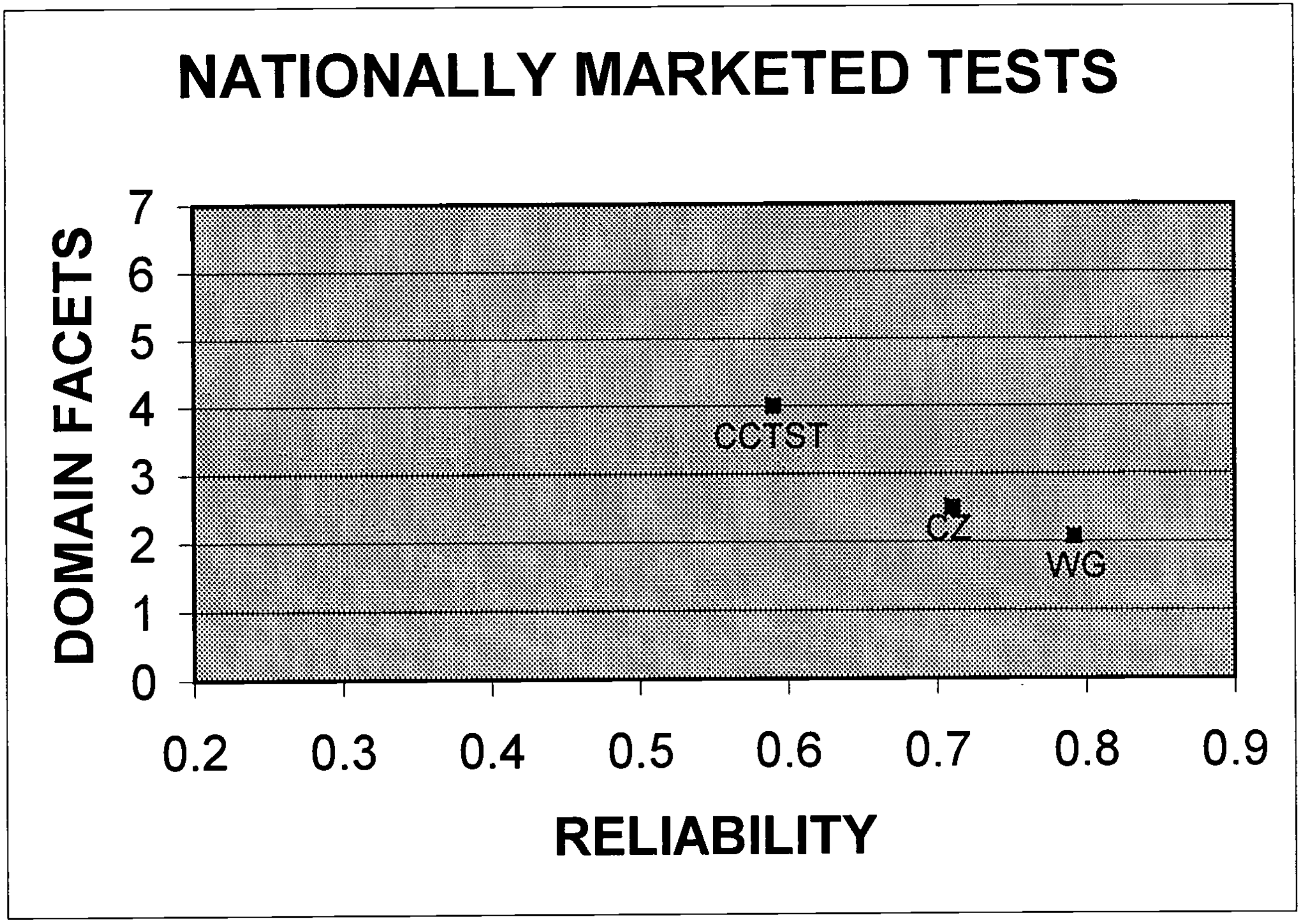
Significance

Although students come to college already able to think, one of the primary purposes of the college experience is to teach them to think even better. This often encompasses providing students with cultural tools that, in application, permit them to make better decisions; that is, help them to become better critical thinkers. Business leaders and academicians alike have identified critical thinking skills as one of the primary skill areas that are both desirable and transferable. Unfortunately, our ability to measure critical thinking in students has not been on par with our desire to teach it. The research presented here proposes an alternative approach, using multiple and mixed methodologies, to measure student critical thinking progress or lack thereof.

Despite trends in business and higher education toward stressing higher order thinking skills over content (Chronicle of Higher Education, June 14, 2000), there currently exist no reliable nationally available instruments that measure the seven facets of critical thinking used in this study: analysis, evaluation, interpretation, inference, dispositions, metacognition/self-regulation, and presenting arguments. In addition, none of the available instruments consistently measure critical thinking outside of a specific disciplinary context. There is a propensity for the nationally marketed CT tests to measure only a narrow portion of the domain and there exists an inverse relationship between domain coverage and reliability (See Figure 1.)

While each of the tests plotted in Figure 1 have definite strengths, from a psychometric perspective they also have weaknesses. Although the California Critical Thinking Skills Test (CCTST) covers more of the domain than the other instruments reviewed here, Cronbach alpha estimates of its reliability are consistently near .58 (Erwin, 2001). Scores on the Cornell Z (CZ) are typically more reliable than the CCTST (Cronbach's alpha=.68-.72; Erwin, 2001), but it thoroughly addresses only 2 facets of the CT domain (analysis and evaluation). The Watson Glaser (WG), while yielding the most reliable test scores, covers the least amount of the critical thinking domain of the tests reviewed here.

Figure 1. Comparison of Nationally Marketed Critical Thinking Tests



Methods

Sample

We report results for both a pilot study and a subsequent study of the revised instrument. The pilot sample consisted of 750 incoming freshmen participating in James Madison University's (JMU) Fall 2000 Assessment Days (August 25-26, 2000). The sample for the subsequent study consisted of 730 sophomores participating in JMU's Spring 2001 Assessment Day (February 22, 2001).

Measures

Critical Thinking Assessment (CTA). This novel multiple-choice critical thinking instrument is a product of more than a year of weekly sessions focusing on the understanding and measurement of critical thinking skills. A number of existing CT instruments were reviewed and the advice of critical thinking experts and measurement experts was incorporated. Additional item level information was gleaned from existing empirical data. The CTA was designed to measure 4 skill areas: analysis, evaluation, inference, and interpretation.

In the pilot study, 2 forms of the skills instrument were administered to maximize the number of items for which information was gathered (399 students took Form A and 251 took Form B). Classical test theory item analysis revealed a small subset of items that were either negatively correlated or not at all correlated with total scores on the CT skills measure. A new form of the assessment consisting of the better items in each skill area across the 2 test forms was compiled (the CTA) and administered to the second sample (N=543). The test took students approximately 30 minutes to complete. There are a total of 40 items on the CTA with 2-5 options each. Here is a sample question from the measure:

"**Statement 1:** College GPAs have increased every year for the past 10 years.

<u>Conclusion</u>: This is a demonstration of grade inflation (the achievement of higher grades than was actually deserved).

Which, if any, of the following statements are <u>unstated</u> assumptions of the speaker?

College students tend to study more than they did over the previous 10 years.
A. An unstated assumption.
B. <u>Not</u> an unstated assumption."

<u>Dispositions (AOT).</u> This 38-item likert-scaled dispositions inventory is a shortened version of the full AOT compiled and administered by Stanovich & West (1997). It was designed to measure the dispositional facet of the critical thinking domain. We sampled from the longer version based upon empirical results (West, personal communication, May, 2000). The shortened scale is comprised of subsets of seven other measures: the Flexible Thinking Scale (Stanovich & West, 1997), NEO Personality Inventory – Openness-Ideas Subscale (Costa & McCrae, 1992), NEO Personality Inventory – Openness-Values Subscale (Costa & McCrae, 1992), Need for Cognition (Cacioppo, et al., 1996), Dogmatism (Rokeach, 1960 as reported in Trohdahl & Powell, 1965; Paulhus & Reid, 1991; and Robinson, Shaver, & Wrightsman, 1991), Absolutism – Subscale of the SID (Erwin, 1981, 1983), and Categorical Thinking (Epstein & Meier, 1989).

In the Fall 2000 pilot, 750 students took the shortened AOT. In the subsequent testing 343 students were administered the shortened AOT. This measure required approximately 10 minutes to complete. Students read individual statements and responded on a 6-point Likert scale: Agree Strongly, Agree Moderately, Agree Slightly, Disagree Slightly, Disagree Moderately, Disagree Strongly. Here is a sample item from the measure:

"Changing your mind is a sign of weakness."

Ways We Think (WWT) Metacognition. The original 27-item Ways We Think is comprised of 3 subscales targeting 3 proposed areas of metacognition. The three general components of this model are an awareness of one's thinking process, an evaluation of one's thinking process, and any necessary or desired adjustments. We therefore included self-awareness, self-monitoring, self-examination, and self-correction (Flavell, 1979; Pintrich, 2000; Delphi Report, 1990). We deleted 4 items that showed near zero correlations with the total score. The resulting subset of 23 items was used for further analysis.

The results presented here represent our pilot study of this instrument. It was administered to 196 sophomores on Assessment Day, February 22, 2001. The measure requires approximately 5 minutes to complete. Students responded on a 5-point frequency scale that expressed how often he or she engaged in the behavior described: Never true of me, Rarely true of me, Sometimes true of me, Frequently true of me, Always true of me. Here is a sample item from the measure:

"When I encounter problems that I cannot solve, I seek more information."

Writing Rubric (CTWRITE). A rating scale was developed that addresses the presenting arguments facet of the critical thinking domain (e.g. the presence of a conclusion, soundness of an argument, relevance of evidence, credibility of sources). We currently possess a sample of student on-demand literature reviews, but have not yet applied the rubric to that sample. As a result this section is not yet complete.

Using generalizability theory, we will study the sources of variability and the generalizability of the novel rubric. We will use a fully crossed 2-facet design with at least 2 raters. We will randomly select 30 writing samples to be used for the initial rubric study. We also intend to modify the CTWRITE rubric for application to student speeches. The CTWRITE rubric consists of 18 criteria to which the rater reports the effectiveness with which students address each criterion. Raters respond on a 5-point scale ranging from 0 (not effective) to 4 (fully effective). Here is a sample item from the measure:

"The conclusion is clearly stated."

Procedures

On Spring Assessment Day, subgroups of the 730 2nd semester sophomores were administered the Critical Thinking Assessment (CTA), AOT Dispositions, Ways We Think, Cornell Z, and California Critical Thinking Dispositions Inventory. For each of these measures higher scores are operationalized as having more of the construct being measured. For instance, higher scores on the CTA indicate more fully developed critical thinking skills and higher scores on the AOT indicate greater openness to belief change and greater cognitive flexibility.

Analytic Strategies

We used 5 different analytic strategies. For each locally developed measure Classical Test Theory (CTT) analyses and Exploratory Factor Analyses (EFA) were performed. Item Response Theory (IRT) procedures were used to investigate the psychometrics of the skills measure including person and item misfit, information functions, and the standard error of measurement. Differential item functioning (DIF) with respect to gender was investigated using the Mantel-Haenszel procedure. Finally, correlations among and between these measures and external measures (Cornell Z, California Critical Thinking Dispositions Inventory, and Scholastic Achievement Test) were performed.

Results

Classical Test Theory Analysis

Using classical test theory we computed the means, standard deviations, reliability as estimated by Cronbach's alpha, the difficulty or p-values, and the corrected item total point biserial correlations. The relative reliability of the measures used indicated that of the skills measures, scores on the CTA are more reliable than scores on the Cornell Z. (See Table 1.) While each of the disposition measures (Actively Open-Minded Thinking (AOT) and California Critical Thinking Dispositions Inventory (CCTDI)) obtained a fairly stable estimated reliability (.89), the AOT did so with just over half as many items (38 versus 75).

Table 1 – Estimated Reliabilities

Instrument	N	Number of Items	Time to Complete	Reliability
Cornell Z	302	52	50 minutes	.63
CCTDI	586	75	20 minutes	.89
CTA	534	40	30 minutes	.83
AOT Dispositions	343	38	10 minutes	.89
WWT	196	23	5 minutes	.83

Exploratory Factor Analysis (EFA)

EFA's (using principal axis factoring) were performed for each of the locally designed measures (the CTA, AOT, and WWT). Conceptually, for each measure we would expect the individual factors (sub-skills, dispositions, metacognitive areas) to be related to each other, not independent, so to improve the interpretability of the solutions we performed oblique factor rotations.

CTA.

The scree plot indicated that the CTA was best described with a 1- factor solution. A priori we proposed either a 4-factor or a 1-factor solution (the four skills we mapped out or the belief that the factors are so closely inter-related that one factor might stand out even though it is not responsible for a preponderance of the variance).

We further analyzed the factor structure for the CTA with a full information factor analysis that uses tetrachoric rather than point biserial correlations. The advantage obtained by using tetrachoric correlations is correction for the dichotomization of item scores. In other words, this technique projects the relationship between variables as if scoring was continuous rather than dichotomized. It thereby removes the propensity for items with similar difficulty levels (rather than similar content) to group together. The full information factor analysis also suggested a 1-factor solution. For the 4-factor solution, while there were some high loadings, only 25 of 40 items loaded at or above .3 on those 4 factors.

To conclude that a test is reasonably unidimensional for IRT analysis, we require the ratio of first to second factor eigenvalue to be 3 to 1 or greater. For this data set, the first eigenvalue (9.36) was 2.5 times the second eigenvalue (3.72). (See Table 2.) While this ratio is a bit less than what we would ideally like, when we extracted just one factor, it accounted for 23% of the variance. In addition, 29 of 40 items loaded at or above .3 on this single factor. We therefore used a 1-factor solution.

Table 2 – CTA Factor Analysis

Eigenvalue	ApproximateAmount of Variance
1st = 9.36	23%
2nd = 3.72	9.3%
3rd = 2.23	5.6%
4th = 2.13	5.3%

AOT.

Exploratory factor analysis confirmed previous findings (Stanovich & West, 1997) that the AOT is reasonably unidimensional, though in our shorter version, the first factor accounted for less variance (21% here versus 38.7%) than reported previously (Stanovich & West, 1997). Because 4 eigenvalues were greater than one (8.0026, 3.8192, 1.3076, 1.0097), we perused 1-, 2-, 3-, and 4- factor solutions. Both the 3- and 4-factor solutions possessed multiple instances of item overlap between factors and hence, their interpretability was limited. Because the 2nd factor accounted for 10% of the variance and because the 2-factor solution cleanly broke along subscale divisions, the 2-factor solution was favored (see Table 3). The 2 factors could be characterized as the propensity to think in a more open-minded manner (subscales loading are Flexible Thinking, Openness-Values, Dogmatisim, Absolutism, and Categorical Thinking) and motivation for cognitive complexity (subscales loading are Openness-Ideas, Need for Cognition). Of the 38 total items, only 4 did not load on one of the 2 factors at .30 or greater and 34 of the 38 items loaded at .32 to .69 on one of the two factors.

Table 3 – AOT Factor Analysis

Eigenvalue	Amount of Variance
1st = 8.0026	21%
2nd = 3.8192	10%
3rd = 1.3076	3.4%
4th = 1.0097	2.7%

Ways We Think.

EFA indicated that the WWT was reasonably unidimensional. Because 4 eigenvalues were greater than one we could interpret either a 1-, 2-, 3-, or 4-factor solution. The scree plot indicated that a 1- or 2-factor solution might be most interpretable (see Table 4). Despite the fall off in variance accounted for, a review of item loadings indicated that if a 1-factor solution was used, some items would not load not factor one and would be lost. The 4-factor solution contained multiple overlap and loaded in an unpredictable fashion. A 2-factor solution provided a nice compromise with all of the items loading on one of the two factors. There were only 5 cases of overlap with the 2-factor solution. Conceptually the 2 factors appeared to have items with a different focus: self-directed seeking of information when needed versus a more persuadable, social orientation.

Table 4 – WWT Factor Analysis

Eigenvalue	Amount of Variance
1st = 4.7121	21%
2nd = 1.4396	6.5%
3rd = 1.129	5%
4th = 1.0489	4.8%

Item Response Theory (IRT) Applied to the CTA

We used Bilog-MG (Zimowsky, Muraki, Mislevy, and Bock, 1996), so parameter estimation procedures were carried out using a marginal maximum likelihood (MML) approach. A modified 2-Parameter Logistic Model (2-PL) in which the guessing parameter was fixed at .122 was used to calibrate items responses on the CTA. We originally set the guessing parameter at .1, but Bilog-MG modified 2-PL calibration indicated that .122 was better suited to the data. Because of the small sample size (N=534), a 3-PL model failed to converge. The 2-PL model provided an estimated mean threshold (difficulty or b-parameter) of -.737 and an average slope of .923. While an N of at least 1,000 is desirable for stable results using MML (Hambleton & Swaminathan, 1991) for a modified 2-PL (with fixed c) or 3-PL estimation, the modified 2-PL calibration performed here provided reasonably stable results.

Table 2 – Summary of Parameter Estimates 2-PL with Fixed c

Parameter	Mean	SD
Person Ability	0.000	1.000
Item Difficulty	-0.737	1.625
Item Discrimination	0.923	.0.793
Guessing Parameter	0.122	0.001

The range of item difficulties was large (-4.96 to 2.34). The overall mean threshold (-0.737) implies that the items on this test were on the easy side for these students (see Table 5). The standard errors about the estimates were greater than ideal indicating, in part, that a larger sample might be beneficial.

We performed analyses regarding IRT assumptions. IRT assumptions are more stringent than CTT. In particular, the data is assumed to be unidimensional, that is, dominated by one factor and each item is assumed to be locally independent. We also assume that the test was not speeded.

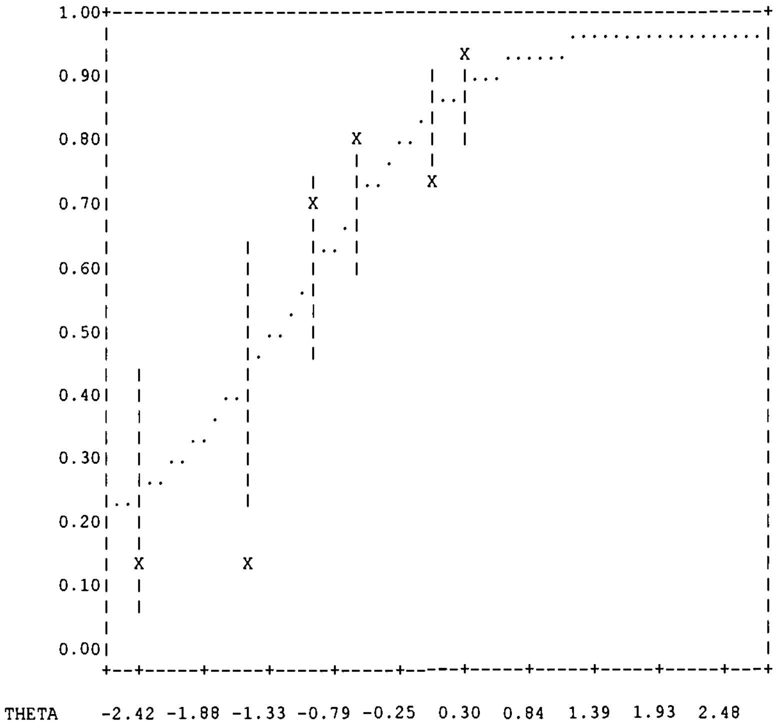
The assumption of unidimensionality, as previously indicated, was queried using Testfact (Scientific Software, Inc., 1985) to run a full information factor analysis. We concluded that the test is reasonably unidimensional. Regarding the assumption of local independence, a review of item content indicated that correctly answering any one item on the test should not affect the probability of correctly answering another test item. In consideration of possible effects of speededness, a review of item data revealed that almost all students completed the test. Missing data was scarce and was as likely to appear in the mid-section as near the end of the test.

To summarize, the assumptions underlying IRT analysis were generally met. The evidence generated to address assumptions of unidimensionality was the most problematic. This does not necessarily invalidate IRT procedures for analyzing the performance of the CTA, but it does give one pause. To glean a clearer picture of test performance and model match, we focus next upon item and person fit.

Item fit was queried using the modified 2-PL calibration. The first step was to look at the chi-square values and associated degrees of freedom for each item estimate in the Bilog-MG output. We used a ratio of chi-square to degrees of freedom equal to 4:1 as the criterion for deciding whether an item was misfit by this method. Using this approach, 3 items were identified as misfit (Items 2, 3, and 5). Plots of these items appear in Figure 2. Each of these items is a counter factual syllogism and in each case, it appears that examines may be getting answers correct for the wrong reasons.

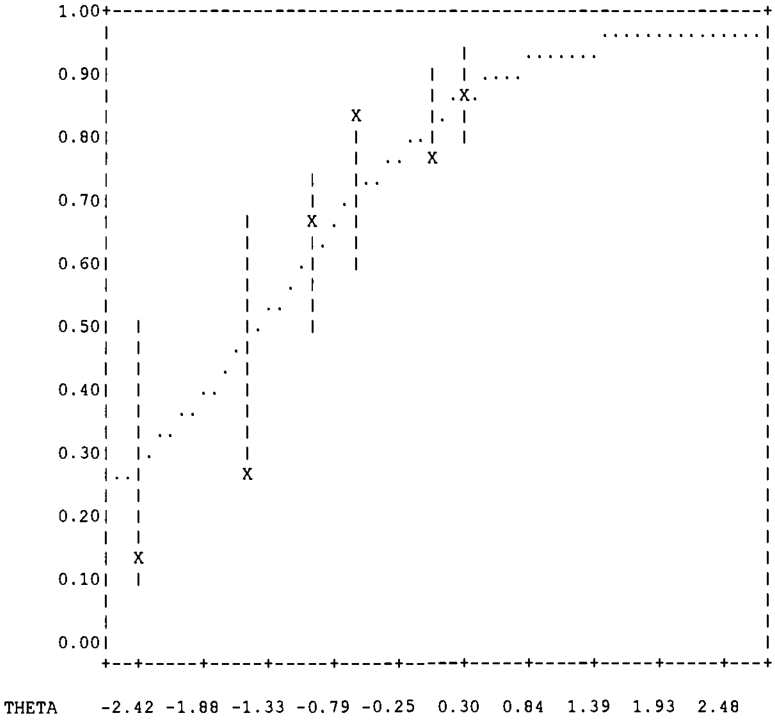
Figure 2 – Misfit Items

Item 2 PROB< 0.0000

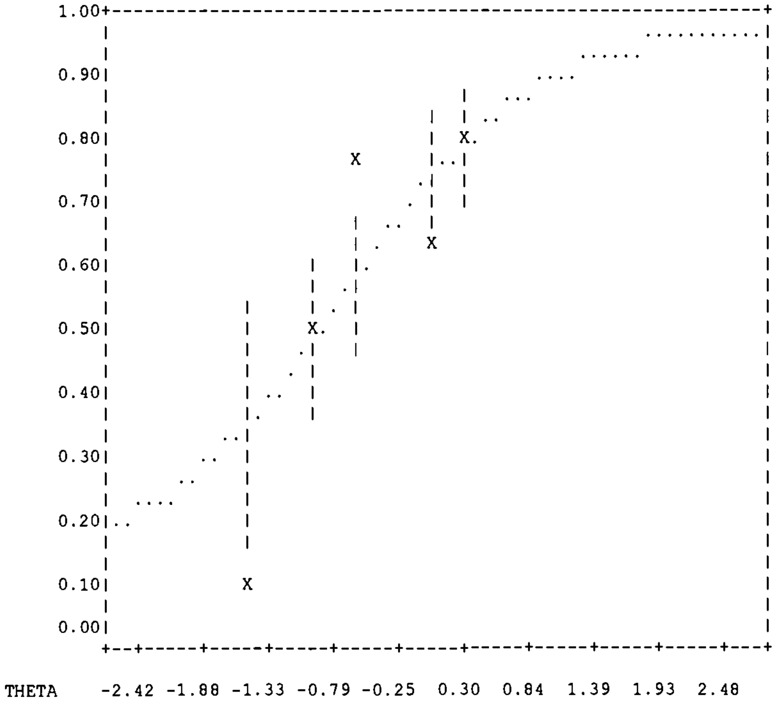


<u>Figure 2 – Misfit Items</u> (continued)

Item 3 PROB< 0.0002



Item 5 PROB< 0.0000



In each of the plots, empirical group scores (represented by the X's) fell outside of the confidence intervals surrounding the ICC's. For Items 2, 3, and 5, the empirical distribution peaked in the mid-range, fell off and peaked again at the highest ability levels. Based upon this analysis, a closer look at these items is warranted. It should be noted; however, that a finding of 3 misfit items out of 40 total items did not greatly exceed the level of chance at a .05 probability for discovering misfit items.

We investigated person-fit using the Appropriateness (lz index) proposed by Levine and Drasgow (1982). After standardizing the person scores we identified those scores whose lz was less than –2.58. This criterion did not identify any response patterns as misfit. We further investigated person fit by comparing the percent correct score (CTT's p-value) and theta averaged across each of 10 groups of examinees. While the pattern generally revealed that as theta or demonstrated proficiency increased so did the percentage correct, it also indicated that this test was not very precise at the lower theta or proficiency levels. From the perspective of the test as whole, the model fit the middle range examinees best.

Information Function.

We focus on test information rather than item information. The test information function is simply the sum of the individual item information functions. Figures 3 and 4 plot the test information function and standard error of measurement (SEM), respectively. On the x-axis is the plotted demonstrated proficiency level from –4 to +4 and on the y-axis is the information or SEM. As evidenced by the mean threshold for the modified 2-PL calibration, the test information function reveals that the CTA provides the greatest information for estimates of theta between –1 and 0. Also obviated by the information plot is the fact that the test does not discriminate as well at theta levels outside the range of –1 to 1. That is, the information is very peaked indicating that there is a steep drop off in precision on either side of this point.

Figure 3 – Test Information Function

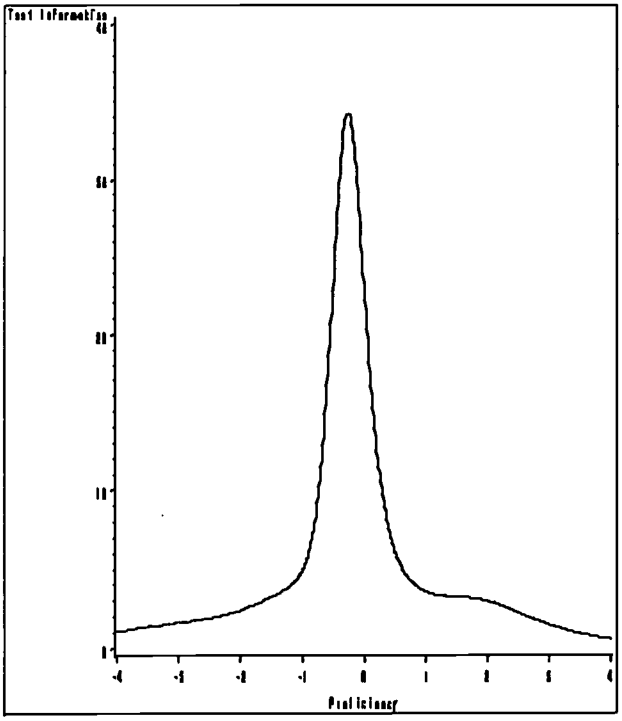
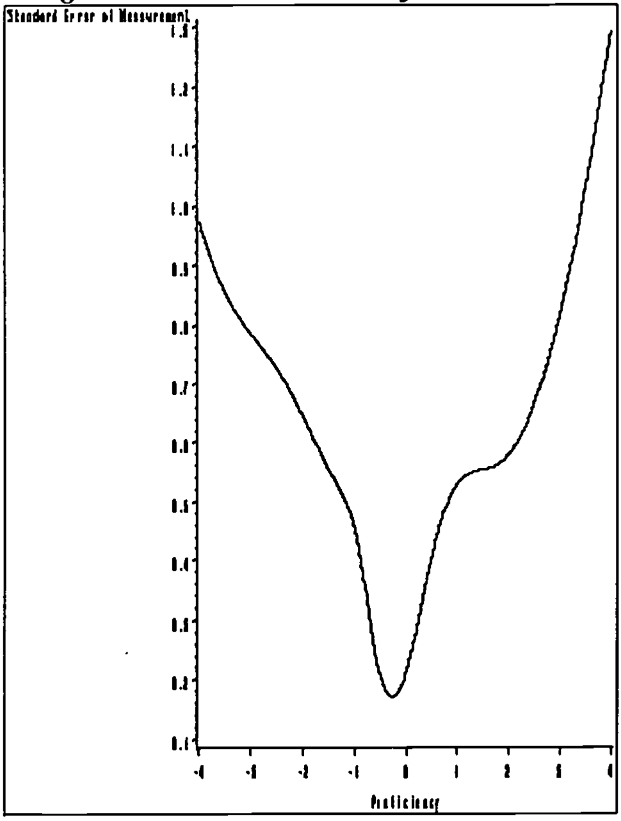


Figure 4 – Standard Error of Measurement



Differential Item Functioning

Inquiry into possible gender DIF was performed using the Mantel-Haenszel (MH) procedure because it is the most stable for smaller groups. Because not all students reported gender, we had less than 200 students per group. The significance level chosen for this analysis was .01 and thus the critical value of chi-square was equal to 6.635. The MH approach revealed 3 items with significant DIF. Items 23, 10, and 13 demonstrated DIF. All 3 deltas are classified as "c's" (>1.5) indicating severe DIF and therefore require closer analysis. At a probability level of .05 though, finding 3 items with DIF out of 40 total items is not appreciably greater than chance. Items 23 and 13 favored women and item 10 favored men. A review of item content revealed plausible reasons for item DIF for Items 10 and 23 (hunting and baseball), but not for Item 13.

Correlations

Pearson's correlations and disattenuated correlations between the different measures are found in Table 8. Pearson's correlations represent the linear relationship between 2 variables. As shown in Table 6, significant pairwise relationships exist between the CTA, AOT, Cornell Z, CCTDI, and SAT. In Table 6, the upper portion of each cell contains the Pearson's correlation coefficient and redundant cells are left blank. The left hand portion of the table contains relationships between each of the measures and the right hand portion reveals relationships between the SAT, CTA, and AOT.

The disattentuated correlations are estimates of the true correlations between the 2 variables after correcting for the variables' reliabilities. In each case, a substantial amount of variance remains unaccounted for by the other variable. Because this correlation is computed using the square root of the product of alpha internal consistency reliabilities for 2 measures, the coefficient appearing here represents an upper limit of the relationship between the 2 variables (alpha underestimates reliability). In Table 6, the disattentuated correlations are italicized and appear in the lower portion of each cell.

The CTA shares common variance with the Cornell Z, AOT, and CCTDI. (Note--Despite the finding that the AOT was best described by a 2-factor solution, we had no theoretical reason to split the score and therefore used only the AOT total score for analyses.) The relationship between the CTA and the Cornell Z (Pearson's r=.58, disattentuated r=.80) was noteworthy because we believed the Cornell Z was measuring a subset of the critical thinking domain (analysis and evaluation) that is measured by the CTA. Thus, we would expect a medium positive correlation between the two; as scores on the Cornell Z increase, scores on the CTA increase. The relationship between the CTA and CCTDI (Pearson's r=.31, disattenuated r=.36) indicated a weak and positive relationship between the two variables. Finally, of particular interest was the medium correlation of the AOT dispositions and the CTA (Pearson's r=.49, disattentuated r=.57). This indicated that a more flexible open minded thinking style was associated with successful application of CT skills.

Other significant relationships between the administered tests included pairwise relationships revealed for both the AOT and Cornell Z and the AOT and CCTDI. A medium, positive relationship was evidenced between the AOT and Cornell Z (Pearson's r=.55, disattenuated r=.73). This was not surprising given the strength of the relationships between the CTA, Cornell Z, and AOT. We also expected and realized a medium positive linear relationship between the AOT and CCTDI (Pearson's r=.59. disattentuated r=.66)

Using SAT as a rough estimate of ability, we also explored the relationships between SAT-Total, SAT-Verbal, and SAT-Math scores and total scores on the CTA and AOT. Not surprisingly, we discovered a medium positive association between the SAT-Total and the CTA (Pearson's r=.45, disattenuated r=.52). This indicated that while the CTA measured some of the same facets SAT measured, it was also substantially singular. The AOT shares its strongest relationship with the SAT-Verbal score (Pearson's r=.37, disattenuated r=.46). This was consistent with previous research (Stanovich & West, 1997). Thus we concluded that the CTA, AOT, and SAT were each assessing some common and some unique domains.

Table 6 – Correlations and Disattentuated Correlations

Instrument	**CTA Skills**	**Cornell Z**	**AOT (38 item)**	**SAT Total**	**SAT Verbal**	**SAT Math**
CTA Skills				.45 n=230 *.52*	.49 n=230 *.57*	.38 n=230 *.44*
Cornell-Z	.58 n=98 *.80*			NA n=0	NA n=0	NA n=0
AOT (38 item)	.49 n=343 *.57*	.55 n=20 *.73*		.31 n=230 *.39*	.37 n=230 *.46*	.17 n=230 *.21*
California Critical Thinking Dispositions	.31 n=319 *.36*	NA N=0	.59 n=219 *.66*	NA n=0	NA n=0	NA n=0

Note—SAT disattentuated correlations are approximate (we used reliability=.90 for SAT). Pearson's correlations appear in the upper portion of each cell and disattenuated correlations are italicized and appear in the lower portion of each cell.

Conclusions & Discussion

The major limitation of this study is that it is not completed. We report a portion of the story, but are unable to present our conceived complete story. Notably missing is within subjects comparison on the CTA, AOT, and WWT. The WWT was given to a separate sample of students (N=196) than the AOT and CTA (N=343). Thus we have no cross comparisons of metacognitive, disposition, and skills measures. Also notably missing are the results of applying a CT rubric to writing and/or speech samples. An initial g-study of the rubric will be performed shortly and in the Fall of 2001 we plan to collect within subjects data on all 4 measures (CTA, AOT, WWT, CTWRITE and or Oral Argument Presentation).

Other limitations are that this sample may not generalize beyond the James Madison University sample. Assessment Days at JMU occur in a low stakes environment (consequences for students are minimal). Corroborative studies both at other universities and in other settings could address this limitation.

Despite limitations, the data presented here indicate that the CTA (skills), AOT (dispositions), and WWT (metacognition/self-regulation) measures demonstrate adequate reliability for making group-level inferences. The psychometric properties of these instruments demonstrate that they function as well, if not better, than other instruments measuring a similar domain. Most strikingly, the CTA, AOT, and WWT administered together measure all but 1 of the seven domain facets of our working definition of critical thinking with an estimated reliability greater than that of any nationally available skills instrument. This is accomplished in approximately 45 minutes.

From a psychometric perspective, there remains more revision on each of the instruments. To achieve reliability in the .90's for the CTA and WWT will require modifying some of the currently used items and perhaps lengthening the tests. With respect to the CTA, we are keenly interested in expanding the information it provides across ability levels. Specifically, we would like more precision in the range of demonstrated proficiency from theta of –2 to 2. While we are presently reasonably accurate from –1 to 1, more items targeted for proficiency levels from 1 to 2 and –1 to –2 are needed. In addition, those items demonstrating gender DIF should be studied further. Due to lack of groups of minority subjects, exploration of other types of DIF were not performed.

The relationships between the novel measures presented here (CTA, AOT, and WWT) and the SAT, Cornell Z, and CCTDI begin the process of assessing the validity of the scores yielded by our measures. While our supposition that good critical thinking skills are not simply a reflection of superior ability was supported by the significant medium (not strong) positive relationship between the CTA and SAT, we clearly need to gather more validity evidence. This effort was begun, in part, to answer the validity concerns of JMU faculty regarding the match between nationally available instruments and the JMU curriculum. Validity studies designed to predict success in chosen major at JMU, graduate school success, and meaningful change over time (throughout the college career and beyond) using CTA, AOT, and WWT scores as predictors should be performed. In general, studies that answer the question, "What does it mean to score high on the CTA, AOT, and WWT in terms of collegiate and/or life success?" need to be undertaken next.

It is our hope that when all the pieces of data are collected within subjects, a Structural Equation Modeling (SEM) approach may elucidate underlying patterns and relationships not currently manifest. For instance, what are the individual differences in critical thinking? Do some domain facets influence each other more than they influence others? Does this occur within the individual? Cognitive modeling approaches to human thought may be ideally suited to this type of inquiry. If we can understand the relationships between the host of skills that comprise critical thinking we can enhance our ability to both teach and measure those skills and perhaps extend that design to other higher order thinking processes.

Acknowledgement

We would like to thank members of our critical thinking research group with whom we meet weekly to discuss the Critical thinking validity and measurement issues. The contributions of Donna Coffman, Keston Fulcher, and Vladislav Tsyganvo have been invaluable.

References

Baron, J. (1998). Judgment misguided: Intuition and error in public decision making. New York, NY: Oxford University Press, Inc.

Baron, J. & Hershey, J.C. (1988). Outcome bias in decision evaluation. Journal of Personality and Social Psychology, 54, 569-579.

Chaffee, J. (1991). Thinking critically, 3rd Edition. Boston, MA: Houghton Mifflin Company.

Ennis, R.H., Millman, J., & Tomko, T.N. (1985). Cornell critical thinking tests level X and level Z: Manual, 3rd Edition. Midwest Publications: Pacific Grove, CA.

Erwin , T.D. (2001). National Postsecondary Educational Cooperative Sourcebook of Assessment Information. Available: http://nces.ed.gov/npec/evaltests/CTandPSInfo.htm.

Facione, P. A. (1990). Executive summary of critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. "The Delphi Report." Milbrae, CA: The California Academic Press.

Flavell, J.H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34, 906-911.

Halpern, D.F. (1997). Critical thinking across the curriculum. A brief edition of thought and knowledge. Mahwah, NJ: Lawrence Earlbaum Associates, Inc.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications, Inc.

Hebel, S. (2000). U.S. workers value general skills over specific ones, survey on education finds. The Chronicle of Higher Education. June 14, 2000.

Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions. Psychological Review, 103, 582-591.

Kelly, D. (1988). The art of reasoning. Markham, Ontario: Penguin Books Canada, Ltd.

Kuhn, D. (1999). A developmental model of critical thinking. Educational Researcher, 28(2), 16-26.

Kuhn, D. (2000). Metacognitive development. Current Directions in Psychological Science. Vol 9(5), 178-181.

Levine, M.V., & Drasgow, F., (1982). Appropriateness measurement: Review, critique and validating studies, British Journal of Mathematical and Statistical Psychology, 35, 42-56.

Manktelow, K. (1999). Reasoning and thinking. Hove, England UK: Psychology Press/Taylor & Francis.

Piattelli-Palmarini, M. (1994). Inevitable illusions: How mistakes of reason rule our minds. United States: John Wiley & Sons, Inc.

Pintrich, P.R. & DeGroot, E.V. (1990). Motivational and self-regulated learning components of classroom academic performance. Journal of Educational Psychology, 82, 33-40.

Pintrich, P.R., Smith, D.A., Garcia, T., and McKeachie, W. (1993), Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). Educational and Psychological Measurement, 53, 801-813.

Pintrich, P.R., Wolters, C.A., and Baxter, G.P. (2000). Assessing metacognition and self-regulated learning.

Ritov, I. & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. Journal of Behavioral Decision Making, 3, 263-277.

Shafir, E., Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. Cognitive Psychology, 24, 449-474.

Stanovich, K.E. & West, R.F. (2000). Individual differences in reasoning: Implications for the rationality debate? Behavioral and Brain Sciences, 23, 645-726.

Stanovich, K.E.,West, R.F. (1999). Discrepancies between normative and descriptive models of decision making and the understanding/acceptance principle. Cognitive Psychology, 38, 1999, 349-385.

Stanovich, K.E, West, R,F. (1999a). Individual differences in reasoning and the heuristics and biases debate. In Ackerman, P.L. and Kyllonen, P.C. (Eds). Learning and individual differences: Process, trait, and content determinants (p.389-411). Washington, DC, USA: American Psychological Association.

Stanovich, K.E., West, R,F. (1998). Individual differences in rational thought. Journal of Experimental Psychology: General, 127(2), 161-188.

Stanovich, K.E, West, R.F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. Journal of Educational Psychology, 89, 342-357.

Testfact 2.12.1 [Computer Software]. (1985). Chicago, IL: Scientific Software, Inc.

Tversky, A., Kahneman, D. (1990). Judgment under uncertainty: Heuristics and biases. In Moser, P.K. (Ed) Rationality in action: Contemporary approaches (p. 171-188). New York, NY, USA: Cambridge University Press.

Watson, G. & Glaser, E.M. (1980). Watson-Glaser critical thinking appraisal: Manual. United States: Harcourt Brace Janovich, Inc.

Zimowsky, Muraki, Mislevy, R.J., Bock, R.D. (1996). Bilog-MG. Scientific Software International, Inc. Chicago, IL

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM033214

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Multifaceted Measurement of Critical Thinking In College Students	
Author(s): Kelly A. Williams, Richard F. West, Steven L. Wise	
Corporate Source: James Madison University	Publication Date: 4-12-01

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☒	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Kelly Williams	Printed Name/Position/Title: Kelly Williams / Graduate Student	
Organization/Address: James Madison University Center for Assessment, MSC 6806 Harrisonburg, VA 22807	Telephone: (540) 568-7169	FAX: (540) 568-7878
	E-Mail Address: willi2ka@jmu.edu	Date: 7/16/01

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com